Google Colab Notebook link:
https://colab.research.google.com/drive/19_fQ6Q7s6xlXpDXOQuQD9tnixh885ysD?usp=sharing

Homework 1 (Total 100 points)

Q1. Load and examine the `Auto.csv` dataset from the course folder on Google drive. **(10 points total)**

1. Should you drop any variable from regression analysis and why? (5 points)
2. Which variables should be treated as numeric and which as categorical? Explain why. (5 points)

Provide python code and analysis results first. Use them to support your answers to the two questions above.

FYI the column definitions (from https://cran.r-project.org/web/packages/ISLR/ISLR.pdf):

- mpg: miles per gallon (The outcome, or y, variable)
- cylinders: Number of cylinders between 4 and 8
- displacement: Engine displacement (cu. inches)
- horsepower: Engine horsepower
- weight: Vehicle weight (lbs.)
- acceleration: Time to accelerate from 0 to 60 mph (sec.)
- year: Model year (modulo 100)
- origin: Origin of car (1. American, 2. European, 3. Japanese)
- name: Vehicle name

```
# Importing required packages
import pandas as pd
from google.colab import drive
import seaborn as sns

# Importing dataset from Drive
drive.mount('/content/drive')
data_folder =
'drive/Othercomputers/asus/MSBA/Fall/BA810Fall23Material/Slides/Data/'
pd.options.mode.chained_assignment = None

auto = pd.read_csv(data_folder+'Auto.csv')
display(auto.head())

Mounted at /content/drive

    mpg  cylinders  displacement horsepower  weight  acceleration
year  \
0  18.0          8         307.0        130    3504          12.0
70
```

```
1  15.0            8          350.0         165     3693             11.5
70
2  18.0            8          318.0         150     3436             11.0
70
3  16.0            8          304.0         150     3433             12.0
70
4  17.0            8          302.0         140     3449             10.5
70

   origin                       name
0       1  chevrolet chevelle malibu
1       1          buick skylark 320
2       1         plymouth satellite
3       1              amc rebel sst
4       1                ford torino
```

```
# Checking number of unique names (i.e. unique car models)
print(auto.name.nunique())

#Dropping name variable
auto = auto.drop("name", axis=1)

304
```

Q1.1 We drop the name column as there are over 300 unique car names, making it infeasible to include them as an encoded categorical variable in our regression model.

```
# Checking data types
display(auto.info())
display(auto.horsepower.unique())

# Dropping rows with missing horsepower values
mask = auto["horsepower"] == '?'
auto = auto[~mask]

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 397 entries, 0 to 396
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   mpg           397 non-null    float64
 1   cylinders     397 non-null    int64
 2   displacement  397 non-null    float64
 3   horsepower    397 non-null    object
 4   weight        397 non-null    int64
 5   acceleration  397 non-null    float64
 6   year          397 non-null    int64
 7   origin        397 non-null    int64
```

```
dtypes: float64(3), int64(4), object(1)
memory usage: 24.9+ KB

None

array(['130', '165', '150', '140', '198', '220', '215', '225', '190',
       '170', '160', '95', '97', '85', '88', '46', '87', '90', '113',
       '200', '210', '193', '?', '100', '105', '175', '153', '180',
'110',
       '72', '86', '70', '76', '65', '69', '60', '80', '54', '208',
'155',
       '112', '92', '145', '137', '158', '167', '94', '107', '230',
'49',
       '75', '91', '122', '67', '83', '78', '52', '61', '93', '148',
       '129', '96', '71', '98', '115', '53', '81', '79', '120', '152',
       '102', '108', '68', '58', '149', '89', '63', '48', '66', '139',
       '103', '125', '133', '138', '135', '142', '77', '62', '132',
'84',
       '64', '74', '116', '82'], dtype=object)
```

Rationale

Dropping rows with missing horsepower values, as they cannot be interpreted by the regression model. Dropping instead of imputing, as number of missing values is low (only 5).

```
#Changing Datatypes

# to Categorical
auto['cylinders'] = auto['cylinders'].astype('category');
auto['year'] = auto['year'].astype('category');
auto['origin'] = auto['origin'].astype('category');

# to numeric
auto['horsepower'] = auto['horsepower'].astype('float');
```

Q1.2 Rationale

To categorical

- Cylinders & Year - As these values are discrete categories and the relationship between them and mpg may not simply be linear, we change them to category dtype
- Origin - these values are encoded categories, they represent countries of origin and not numeric values

To numeric

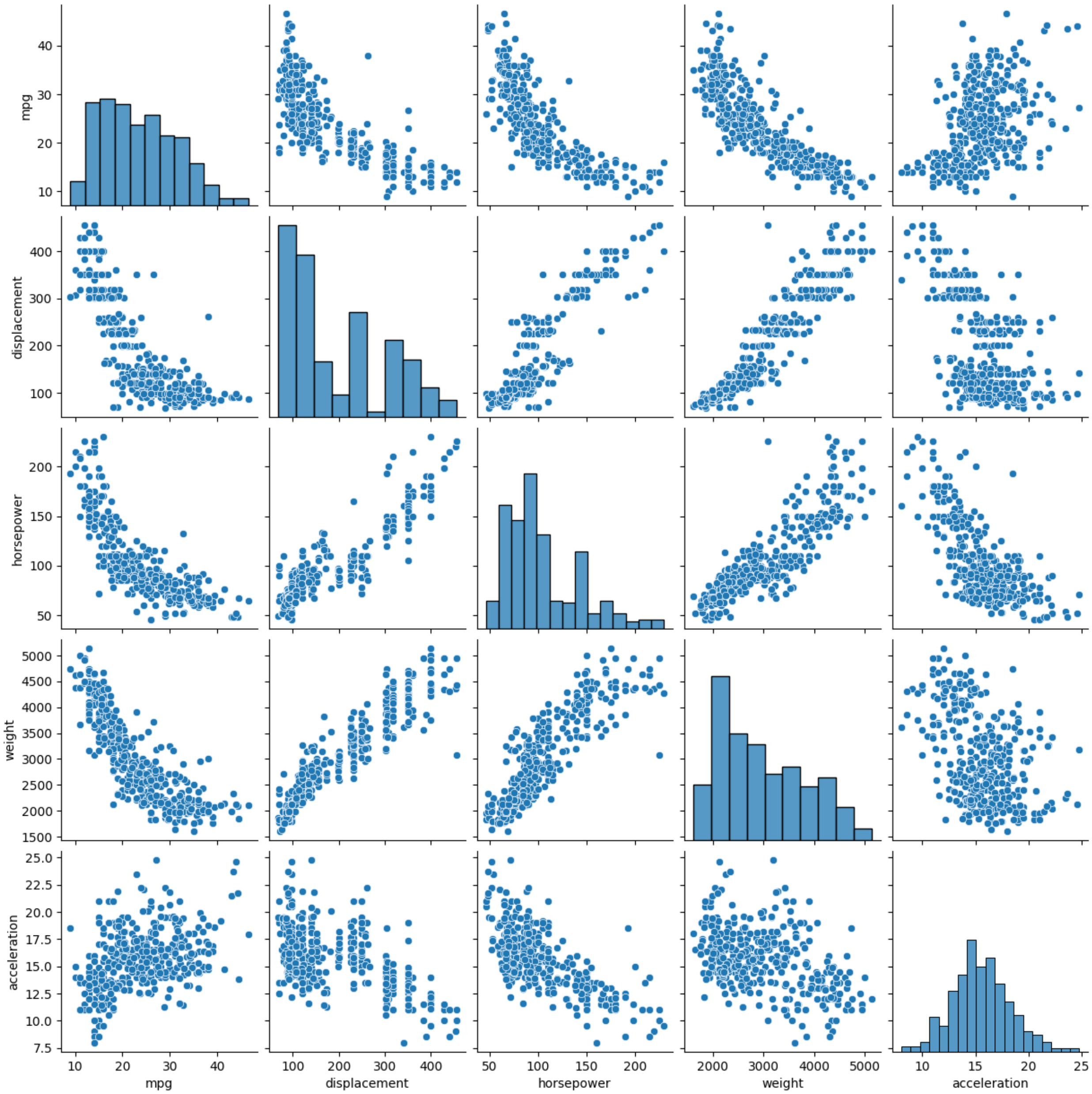
- horsepower - was mistakenly set as object dtype due to `?` missing values.

Q2. Scatter and explore. **(20 points total)**

1. Plot all the pairwise scatter plots and histograms for the numeric features. (10 points)
2. Discuss two interesting relationships that you notice. (10 points)

```
sns.pairplot(auto)
```

```
<seaborn.axisgrid.PairGrid at 0x7e0ad6ee3250>
```


mpg
displacement
horsepower
weight
acceleration

Q2.2 Some interesting relationships

- `displacement`, `horsepower` and `mpg`

  The first two variables describe the "raw power" of the vehicle, and hence have a strong positive correlation with each other. However, the two independently have a strong negative correlation with `mpg`, indicating that the more "powerful" a vehicle is, the less fuel efficient it is likely to be.

- `weight` and `mpg`

  The high negative correlation between vehicle weight and fuel efficiency tells us that lighter weight vehicles tend to be have higher fuel efficiency - something for auto manfacturers to consider during the design process.

Q3. Compute the correlation matrix among the numeric variables. Discuss one interesting correlation. (10+10=**20 points total**)

```
auto.corr(numeric_only =True)
```

```
                   mpg  displacement  horsepower    weight
acceleration
mpg           1.000000     -0.805127   -0.778427 -0.832244
0.423329
displacement -0.805127      1.000000    0.897257  0.932994      -
0.543800
horsepower   -0.778427      0.897257    1.000000  0.864538      -
0.689196
weight       -0.832244      0.932994    0.864538  1.000000      -
0.416839
acceleration  0.423329     -0.543800   -0.689196 -0.416839
1.000000
```

There is a positive correlation between acceleration (time to 60 mph) and mpg (fuel efficiency), one which I did not expect. Conventional wisdom tells us that aggressive driving (quick acceleration and decceleration) reduce fuel efficiency.

However we can see that `acceleration` is negatively correlated to `displacement`, `horsepower` and `weight`, all of which have a negative correlation with `mpg`. This indicates that cars with higher acceleration also tend to have lower values for features that decrease fuel efficiency, resulting in a net positive correlation with `mpg`.

Q4. Use `statsmodels` to regress mpg on all other variables. Note you can tell `ols()` to treat a variable as categorical by enclosing the variable in `C()`. **(15 points total)**

1. Interpret the significant effects. (5 points)
2. Which variables don't have a significant effect? Provide potential explanation for one surprising non-effect. (5 points)
3. Discuss the difference in results when you treat `year` as a categorical vs a numeric variable. (5 points)

```
import statsmodels.formula.api as smf

est = smf.ols('mpg ~
displacement+horsepower+weight+acceleration+cylinders+year+origin',aut
o).fit()
print(est.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    mpg   R-squared:
0.874
Model:                            OLS   Adj. R-squared:
0.867
Method:                 Least Squares   F-statistic:
116.8
Date:                Wed, 01 Nov 2023   Prob (F-statistic):
2.64e-151
Time:                        18:09:00   Log-Likelihood:
-954.59
No. Observations:                 392   AIC:
1955.
Df Residuals:                     369   BIC:
2047.
```

```
Df Model:                          22

Covariance Type:            nonrobust

======================================================================
============
                      coef    std err          t      P>|t|      [0.025
0.975]
----------------------------------------------------------------------
------------
Intercept          30.9168      2.361     13.095      0.000      26.274
35.559
cylinders[T.4]      6.9399      1.537      4.516      0.000       3.918
9.962
cylinders[T.5]      6.6377      2.337      2.840      0.005       2.042
11.234
cylinders[T.6]      4.2973      1.706      2.519      0.012       0.943
7.652
cylinders[T.8]      6.3668      1.969      3.234      0.001       2.495
10.238
year[T.71]          0.9104      0.816      1.116      0.265      -0.693
2.514
year[T.72]         -0.4903      0.804     -0.610      0.542      -2.071
1.090
year[T.73]         -0.5529      0.721     -0.766      0.444      -1.972
0.866
year[T.74]          1.2420      0.855      1.453      0.147      -0.439
2.923
year[T.75]          0.8704      0.837      1.039      0.299      -0.776
2.517
year[T.76]          1.4967      0.802      1.866      0.063      -0.080
3.074
year[T.77]          2.9987      0.820      3.657      0.000       1.386
4.611
year[T.78]          2.9738      0.779      3.816      0.000       1.442
4.506
year[T.79]          4.8962      0.825      5.936      0.000       3.274
6.518
year[T.80]          9.0589      0.875     10.351      0.000       7.338
10.780
year[T.81]          6.4582      0.864      7.477      0.000       4.760
8.157
year[T.82]          7.8376      0.849      9.228      0.000       6.167
9.508
origin[T.2]         1.6933      0.516      3.280      0.001       0.678
2.708
origin[T.3]         2.2929      0.497      4.616      0.000       1.316
3.270
displacement        0.0118      0.007      1.745      0.082      -0.001
```

```
0.025
horsepower        -0.0392      0.013     -3.010      0.003      -0.065
-0.014
weight            -0.0052      0.001     -8.300      0.000      -0.006
-0.004
acceleration       0.0036      0.087      0.042      0.967      -0.167
0.174
======================================================================
========
Omnibus:                       32.560   Durbin-Watson:
1.574
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
55.829
Skew:                           0.528   Prob(JB):
7.53e-13
Kurtosis:                       4.518   Cond. No.
7.95e+04
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 7.95e+04. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

Interpretation

- Q4.1 By looking at the P>|t| values, we can check for the significant variables. Our model has found tha most variables are statistically significant.
- Significant effects: `cylinders`, `origin`, `horsepower`, `weight`. They all have p values <= 0.012, indicating high probability that these variables have a significant effect on mpg.
- Q4.2 `Acceleration`, `displacement`, and certain `years` are not statistically significant (p-value greater than 0.05). Out of these three exceptions, displacement's p-value is only slightly over the 0.05 threshold with 0.082, and all these variables can still be valuable for the model in the task of prediction.
  - ***Surprising non-effect: Acceleration***

    Despite showing a significant correlation with mpg, there is a 96.7% that the effect of this variable is due to random chance. However, we found previously that acceleration is highly correlated with significant predictors such as displacement and weight. This may be the reason for its poor performance here.

```
# using year as numeric variable

auto_num = auto.copy()
auto_num["year"] = auto_num["year"].astype("int")

est_num = smf.ols('mpg ~
displacement+horsepower+weight+acceleration+cylinders+year+origin',aut
o_num).fit()
print(est_num.summary())

                            OLS Regression Results

======================================================================
========
Dep. Variable:                    mpg   R-squared:
0.847
Model:                            OLS   Adj. R-squared:
0.842
Method:                 Least Squares   F-statistic:
191.1
Date:                Wed, 01 Nov 2023   Prob (F-statistic):
2.39e-147
Time:                        18:09:00   Log-Likelihood:
-993.35
No. Observations:                 392   AIC:
2011.
Df Residuals:                     380   BIC:
```

```
2058.
Df Model:                          11

Covariance Type:            nonrobust

==============================================================================
============
                      coef    std err          t      P>|t|      [0.025
0.975]
------------------------------------------------------------------------------
------------
Intercept         -22.0801      4.541     -4.862      0.000     -31.009
-13.151
cylinders[T.4]      6.7218      1.654      4.064      0.000       3.470
9.974
cylinders[T.5]      7.0784      2.516      2.813      0.005       2.131
12.026
cylinders[T.6]      3.3512      1.824      1.837      0.067      -0.236
6.938
cylinders[T.8]      5.0992      2.109      2.418      0.016       0.953
9.246
origin[T.2]         1.7640      0.551      3.200      0.001       0.680
2.848
origin[T.3]         2.6172      0.527      4.964      0.000       1.581
3.654
displacement        0.0187      0.007      2.590      0.010       0.005
0.033
horsepower         -0.0349      0.013     -2.639      0.009      -0.061
-0.009
weight             -0.0058      0.001     -9.154      0.000      -0.007
-0.005
acceleration        0.0260      0.093      0.279      0.780      -0.157
0.209
year                0.7370      0.049     15.064      0.000       0.641
0.833
==============================================================================
========
Omnibus:                       45.781   Durbin-Watson:
1.336
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
85.634
Skew:                           0.677   Prob(JB):
2.54e-19
Kurtosis:                       4.846   Cond. No.
9.32e+04
==============================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
```

```
correctly specified.
[2] The condition number is large, 9.32e+04. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

Q4.3 By using year as a numerical variable instead of a categorical variable:

- Coefficients - year (categorical) has significantly different coefficients for each year, while year (num) has a single coefficient for the variable. This indicates that the model with year (num) is treating the years more uniformly than year(cat).
- Significance - year (num) is statistically significant, but only some years in year (cat) are statistically significant.
- Implications - The year(cat) adjusted R2 is greater than year (num) adjusted R2 - 0.867 vs 0.842. This indicates that year (cat) provides greater explanation of the variance than year (num). The year is more useful as a categorical variable than a numeric one, possibly due to the fact that it has different effects across different years leading to a more nuanced relationship between year and mpg.

Q5. From the above regression model in Q4, include two way interactions between a numeric and categorical variable in three different regression models (three separate models in total). Do any of them appear significant? Discuss the results. **(15 points total)**

horsepower x cylinders

```
est_1 = smf.ols('mpg ~
displacement+horsepower+weight+acceleration+cylinders+year+origin+hors
epower*cylinders',auto).fit()
print(est_1.summary())

                            OLS Regression Results

======================================================================
========
Dep. Variable:                    mpg   R-squared:
0.895
Model:                            OLS   Adj. R-squared:
0.887
Method:                 Least Squares   F-statistic:
119.2
```

```
Date:                  Wed, 01 Nov 2023   Prob (F-statistic):
4.99e-161
Time:                          18:09:00   Log-Likelihood:
-920.12
No. Observations:                   392   AIC:
1894.
Df Residuals:                       365   BIC:
2001.
Df Model:                            26

Covariance Type:              nonrobust

======================================================================
=======================
                                coef    std err          t      P>|t|
[0.025      0.975]
----------------------------------------------------------------------
-----------------------
Intercept                    19.0292     18.494      1.029      0.304
-17.338      55.397
cylinders[T.4]               29.1443     18.506      1.575      0.116
-7.247      65.536
cylinders[T.5]               43.0358     20.392      2.110      0.036
2.935      83.137
cylinders[T.6]               18.4792     18.640      0.991      0.322
-18.175      55.134
cylinders[T.8]               13.8497     18.586      0.745      0.457
-22.700      50.399
year[T.71]                    0.8521      0.761      1.120      0.264
-0.644       2.348
year[T.72]                   -0.1897      0.748     -0.254      0.800
-1.661       1.281
year[T.73]                   -0.5037      0.673     -0.749      0.454
-1.826       0.819
year[T.74]                    0.9611      0.799      1.202      0.230
-0.611       2.533
year[T.75]                    1.1438      0.780      1.467      0.143
-0.390       2.677
year[T.76]                    1.4174      0.748      1.894      0.059
-0.055       2.889
year[T.77]                    2.9563      0.769      3.842      0.000
1.443       4.469
year[T.78]                    3.3278      0.744      4.470      0.000
1.864       4.792
year[T.79]                    5.1270      0.781      6.562      0.000
3.590       6.664
year[T.80]                    8.5395      0.824     10.360      0.000
6.919      10.160
year[T.81]                    6.0493      0.809      7.473      0.000
```

```
4.458       7.641
year[T.82]                        7.8050      0.797      9.797      0.000
6.238       9.372
origin[T.2]                       1.2143      0.479      2.534      0.012
0.272       2.157
origin[T.3]                       1.8367      0.461      3.986      0.000
0.931       2.743
displacement                     -0.0039      0.007     -0.568      0.571
-0.018       0.010
horsepower                        0.0828      0.186      0.445      0.657
-0.283       0.448
horsepower:cylinders[T.4]    -0.2335      0.186     -1.254      0.211
-0.600       0.133
horsepower:cylinders[T.5]    -0.4047      0.212     -1.910      0.057
-0.822       0.012
horsepower:cylinders[T.6]    -0.1325      0.187     -0.709      0.479
-0.500       0.235
horsepower:cylinders[T.8]    -0.0899      0.186     -0.483      0.630
-0.456       0.276
weight                           -0.0033      0.001     -5.311      0.000
-0.005      -0.002
acceleration                     -0.2249      0.085     -2.645      0.009
-0.392      -0.058
======================================================================
========
Omnibus:                       40.702   Durbin-Watson:
1.861
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
89.255
Skew:                           0.555   Prob(JB):
4.15e-20
Kurtosis:                       5.057   Cond. No.
9.74e+05
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 9.74e+05. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

The interaction between horsepower & cylinders is not statistically significant in 4 out of 4 cases - P>|t| ranges from 63% to 5.7%. The general trend seems to be that cylinder size reduces the effect of horsepower on prediction (reduces the coefficient), and the larger the cylinder size the smaller the reduction (with 4 cylinders being the exception to the rule).

The interactions here are valuable to the model despite the large p-values, as shown by the 0.02 increase in adj R2 over the standard model

weight x cylinders

```
est_2 = smf.ols('mpg ~
displacement+horsepower+weight+acceleration+cylinders+year+origin+weig
ht*cylinders',auto).fit()
print(est_2.summary())

                            OLS Regression Results

======================================================================
========
Dep. Variable:                    mpg   R-squared:
0.896
Model:                            OLS   Adj. R-squared:
0.888
Method:                 Least Squares   F-statistic:
120.5
Date:                Wed, 01 Nov 2023   Prob (F-statistic):
9.20e-162
Time:                        18:09:00   Log-Likelihood:
-918.29
No. Observations:                 392   AIC:
1891.
Df Residuals:                     365   BIC:
1998.
Df Model:                          26

Covariance Type:            nonrobust

======================================================================
===================
                            coef    std err          t      P>|t|
[0.025      0.975]
----------------------------------------------------------------------
-------------------
Intercept                19.1911     14.948      1.284      0.200
-10.204      48.586
cylinders[T.4]           28.0128     14.979      1.870      0.062
-1.444      57.469
cylinders[T.5]            9.2018     21.702      0.424      0.672
-33.475      51.878
```

```
cylinders[T.6]              16.3679     15.132      1.082      0.280
-13.389      46.125
cylinders[T.8]               4.2050     15.158      0.277      0.782
-25.603      34.013
year[T.71]                  -0.2712      0.761     -0.356      0.722
-1.767       1.225
year[T.72]                  -0.9551      0.739     -1.292      0.197
-2.409       0.499
year[T.73]                  -1.1719      0.669     -1.753      0.081
-2.487       0.143
year[T.74]                   0.3066      0.794      0.386      0.700
-1.255       1.869
year[T.75]                   0.7351      0.772      0.952      0.342
-0.784       2.254
year[T.76]                   1.1114      0.739      1.503      0.134
-0.343       2.566
year[T.77]                   2.2682      0.762      2.976      0.003
0.770       3.767
year[T.78]                   2.9339      0.721      4.067      0.000
1.515       4.353
year[T.79]                   4.7820      0.758      6.308      0.000
3.291       6.273
year[T.80]                   9.0173      0.805     11.205      0.000
7.435      10.600
year[T.81]                   6.1598      0.793      7.767      0.000
4.600       7.719
year[T.82]                   7.6532      0.779      9.821      0.000
6.121       9.186
origin[T.2]                  1.4016      0.475      2.949      0.003
0.467       2.336
origin[T.3]                  1.3442      0.470      2.861      0.004
0.420       2.268
displacement                 0.0082      0.006      1.302      0.194
-0.004       0.021
horsepower                  -0.0371      0.012     -3.085      0.002
-0.061      -0.013
weight                       0.0004      0.006      0.062      0.950
-0.012       0.013
weight:cylinders[T.4]       -0.0091      0.006     -1.461      0.145
-0.021       0.003
weight:cylinders[T.5]       -0.0023      0.008     -0.289      0.773
-0.018       0.013
weight:cylinders[T.6]       -0.0053      0.006     -0.846      0.398
-0.018       0.007
weight:cylinders[T.8]       -0.0018      0.006     -0.294      0.769
-0.014       0.010
acceleration                -0.0088      0.081     -0.110      0.913
-0.167       0.150
======================================================================
```

```
========
Omnibus:                       31.309   Durbin-Watson:
1.686
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
83.967
Skew:                           0.350   Prob(JB):
5.85e-19
Kurtosis:                       5.157   Cond. No.
9.43e+05
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 9.43e+05. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

There seems to be no benefit to the interaction between weight and cylinders.

- All are statistically insignificant, with p-values from 15% to 77%
- The effect coefficients are small, ranging from -0.002 to -0.009. However, the weight coefficient is equally small (0.0004), and the weight is in pounds (ranging from 1613 lb to 5140 lb). So the net change in mpg prediction is significant.

There is a 0.021 increase in adjusted R2 over baseline which indicates the interaction is useful for prediction

displacement x origin

```
est_3 = smf.ols('mpg ~
displacement+horsepower+weight+acceleration+cylinders+year+origin+disp
lacement*origin',auto).fit()
print(est_3.summary())

                            OLS Regression Results

======================================================================
========
Dep. Variable:                    mpg   R-squared:
0.881
Model:                            OLS   Adj. R-squared:
0.873
Method:                 Least Squares   F-statistic:
113.0
Date:                Wed, 01 Nov 2023   Prob (F-statistic):
```

```
2.15e-153
Time:                            18:09:01   Log-Likelihood:
-944.27
No. Observations:                     392   AIC:
1939.
Df Residuals:                         367   BIC:
2038.
Df Model:                              24

Covariance Type:            nonrobust

======================================================================
======================
                              coef    std err          t      P>|t|
[0.025      0.975]
----------------------------------------------------------------------
----------------------
Intercept                  26.9649      2.471     10.914      0.000
22.106      31.824
cylinders[T.4]              9.6535      1.649      5.853      0.000
6.410      12.897
cylinders[T.5]             11.0373      2.489      4.434      0.000
6.142      15.933
cylinders[T.6]              7.4105      1.843      4.021      0.000
3.786      11.035
cylinders[T.8]              8.4749      2.018      4.200      0.000
4.507      12.443
year[T.71]                  0.6515      0.799      0.816      0.415
-0.919       2.223
year[T.72]                 -0.4649      0.786     -0.592      0.554
-2.010       1.080
year[T.73]                 -0.6107      0.706     -0.865      0.387
-1.999       0.777
year[T.74]                  0.7876      0.842      0.935      0.350
-0.868       2.444
year[T.75]                  0.7531      0.819      0.919      0.359
-0.858       2.364
year[T.76]                  1.5598      0.783      1.991      0.047
0.019       3.100
year[T.77]                  2.9192      0.802      3.638      0.000
1.341       4.497
year[T.78]                  3.0583      0.761      4.017      0.000
1.561       4.555
year[T.79]                  5.0391      0.806      6.250      0.000
3.454       6.625
year[T.80]                  9.1149      0.858     10.623      0.000
7.428      10.802
year[T.81]                  6.7086      0.845      7.935      0.000
5.046       8.371
```

```
year[T.82]                     7.9726      0.831      9.597      0.000
6.339       9.606
origin[T.2]                    8.1298      2.083      3.902      0.000
4.033      12.227
origin[T.3]                    9.0254      1.835      4.920      0.000
5.418      12.633
displacement                   0.0042      0.007      0.616      0.539
-0.009       0.018
displacement:origin[T.2]    -0.0573      0.018     -3.126      0.002
-0.093      -0.021
displacement:origin[T.3]    -0.0613      0.016     -3.725      0.000
-0.094      -0.029
horsepower                  -0.0322      0.013     -2.507      0.013
-0.057      -0.007
weight                      -0.0041      0.001     -6.239      0.000
-0.005      -0.003
acceleration                -0.0750      0.087     -0.859      0.391
-0.247       0.097
======================================================================
========
Omnibus:                       35.321   Durbin-Watson:
1.634
Prob(Omnibus):                  0.000   Jarque-Bera (JB):
69.234
Skew:                           0.521   Prob(JB):
9.25e-16
Kurtosis:                       4.776   Cond. No.
9.06e+04
======================================================================
========

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is
correctly specified.
[2] The condition number is large, 9.06e+04. This might indicate that
there are
strong multicollinearity or other numerical problems.
```

There apprears to be a significant interaction between horsepower & origin:

- Low p-values (nearly 0)
- The effect of origin on displacement is strong and non-zero
- There's a 0.05 increase in adjusted R2 from baseline

Q6. Measure the in-sample and out of sample R^2 of the model specified in Q4.1 using 80% data for training and 20% data for testing. **(10 points total)**

```
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score

auto_train, auto_test = train_test_split(auto, test_size = .20,
random_state=99)
est_final = smf.ols('mpg ~
displacement+horsepower+weight+acceleration+cylinders+year+origin',aut
o_train).fit()
print('in-sample r-square: {:.2f}'.format(est_final.rsquared))

predictions = est_final.predict(auto_test)
print('out-of-sample r-square: {:.2f}'.format(r2_score(auto_test.mpg,
predictions)))

in-sample r-square: 0.88
out-of-sample r-square: 0.83
```

The out-of-sample R2 is lower than in-sample R2, but the difference is to be expected due to generalization error.

Our out-of-sample R2 is large enough to conclude that our model is useful for predicting mpg.

Q7. Collaboration statement **(10 points total)**

I did not get any help from any generative AI tool.

I discussed the homework with Raiymbek Ordabayev, to help him better understand what was required to be done as part of the analysis.